

# Measuring Generalized Preferential Attachment in Dynamic Social Networks

Camille Roth\*

February 3, 2008

## Abstract

The mechanism of preferential attachment underpins most recent social network formation models. Yet few authors attempt to check or quantify assumptions on this mechanism. We call *generalized preferential attachment* any kind of preference to interact with other agents with respect to any node property. We then introduce tools for measuring empirically and characterizing comprehensively such phenomena, and apply these tools to a socio-semantic network of scientific collaborations, investigating in particular homophilic behavior. This opens the way to a whole class of realistic and credible social network morphogenesis models.

*Keywords:* *Morphogenesis models, Preferential attachment, Social Networks, Dynamic Networks, Complex systems.*

ACM: G.2.2; MSC: 68R10; PACS: 89.65.-s, 87.23.Ge, 89.75.-k

## Introduction

A recent challenge in structural research in social science consists in modeling social network formation. Social networks are usually interaction networks — nodes are agents and links between nodes represent interactions between agents — and in this respect, modeling them involves disciplines linked both to graph theory (computer science and statistical physics), mathematical sociology and economics [1, 9, 26]. Most of the interest in this topic stems from the empirical observation that real social networks strongly differ from uniform

random graphs as regards several statistical parameters; and foremost with respect to node connectivity distribution, or *degree* distribution. Indeed, in random graphs *a la* Erdos-Renyi [12] links between agents are present with a constant probability  $p$  and degree distributions follow a Poisson law whereas empirical social networks exhibit power-law, or *scale-free*, degree distributions [1]. This phenomenon suggested that link formation does not occur randomly but instead depends on node and network properties — that is, agents do not interact at random but instead according to heterogeneous preferences for other nodes.

Hence, early social network models endeavored to describe non-uniform interaction and growth mechanisms yielding the famous “scale-free” degree distribution [2]. Subsequently, much work has been focused on determining processes explaining and rebuilding more complex network structures consistent with those observed in the real world — a consistency validated through a rich set of statistical parameters measured on empirical networks, not limited to degree distribution but including as well clustering coefficient, average distance, assortativity, etc. [6, 7, 20].

However, even when cognitively, sociologically or anthropologically credible, most of the hypotheses driving these models are mathematical abstractions that do not enjoy any experimental measurement or justification. In this paper, we call *preferential attachment* (PA) any kind of non-uniform interaction behavior and introduce tools for empirically measuring PA with respect to any node property. We eventually apply these measures to an empirical case of socio-semantic network. In particular, we will criticize degree-related PA, and estimate homophily.

---

\*CREA (Center for Research in Applied Epistemology), CNRS/Ecole Polytechnique, 1 rue Descartes, 75005 Paris, France. E-mail: [roth@shs.polytechnique.fr](mailto:roth@shs.polytechnique.fr)

# 1 A brief survey of social network models

Barabasi & Albert [2] pioneered the use of preferential linking in social network formation models to successfully rebuild a particular statistical parameter, the scale-free degree distribution. In their model, new nodes arrive at a constant rate and attach to already-existing nodes with a likeliness linearly proportional to their degree. This model has been widely spread and reused, and consequently made the term “preferential attachment” often understood as degree-related only preferential attachment. Since then, many authors introduced diverse modes of preferential link creation depending on either various node properties (hidden variables and “types” [5, 28], fitness [7], centrality, euclidian distance [13], common friends [16], bipartite structure [14], etc.) or on various linking mechanisms (competitive trade-off and optimization heuristics [10, 13], two-steps node choice [29], to cite a few).

However and even in recent papers, hypotheses on PA are often arbitrary and at best supported by qualitative intuitions. Existing *quantitative* estimations of PA and consequent validations of modeling assumptions are extremely rare. In this respect, most studies are either (i) related to the classical degree-related PA [3, 11, 15, 23], sometimes extended to a selected network property, like common acquaintances [19]; or (ii) reducing PA to a single parameter: for instance using econometric approaches [21] or Markovian models [27]. While of great interest in approaching the underlying behavioral reality of social networks, these works may not be able to provide a sufficient empirical basis and support for designing trustworthy PA mechanisms, and accordingly for proposing credible social network morphogenesis models. Yet in this view we argue that the three following points are key:

1. Node degree does not make it all — and even the popular degree-related PA (a linear “rich-get-richer” heuristics) seems to be inaccurate for some types of real networks [3], and possibly based on flawed behavioral fundations, as we will suggest below in Sec. 4.2.
2. Strict social network topology and derived properties may not be sufficient to account for complex social phenomena — as several

above-cited models suggest, introducing “external” properties (such as e.g. node types) may influence interaction; explaining for instance homophily-related PA [18] requires at least to *qualify* nodes.

3. Single parameters cannot express the rich heterogeneity of interaction behavior — for instance, when assigning a unique parameter to preferential interaction with close nodes, one misses the fact that such interaction could be significantly more frequent for very close nodes than for loosely close nodes, or be quadratic with respect to the distance, etc.

To summarize, it is thus crucial to conceive PA in such a way that (i) it is a flexible and general mechanism, depending on relevant parameters based on both topological and non-topological properties; and (ii) it is an empirically valid function describing the whole scope of possible interactions.

## 2 Measuring preferential attachment

PA is the likeliness for a link to appear between two nodes with respect to node properties. In order to measure it, we first have to distinguish between (i) single node properties, or *monadic* properties (such as degree, age, etc.) and (ii) node dyad properties, or *dyadic* properties (social distance, dissimilarity, etc.). When dealing with monadic properties indeed, we seek to know the propensity of some kinds of nodes to be involved in an interaction. On the contrary when dealing with dyads, we seek to know the propensity for an interaction to occur preferentially with some kinds of couples.<sup>1</sup>

### 2.1 Monadic PA

Suppose we want to measure the influence on PA of a given monadic property  $m$  taking values in  $\mathcal{M} = \{m_1, \dots, m_n\}$ .<sup>2</sup> We assume this influence can

<sup>1</sup>Note that a couple of monadic properties can be considered dyadic; for instance, a couple of nodes of degrees  $k_1$  and  $k_2$  considered as a dyad  $(k_1, k_2)$ . This makes the former case a refinement, not always possible, of the latter case.

<sup>2</sup>This topic will be described more extensively in a forthcoming general paper presenting basic parameters for dynamic network analysis [17].

be described by a function  $f$  of  $m$ , independent of the distribution of agents of kind  $m$ . Denoting by “L” the event “attachment of a new link”,  $f(m)$  is simply the conditional probability  $P(L|m)$  that an agent of kind  $m$  is involved into an interaction.

Thus, it is  $f(m)$  times more probable that an agent of kind  $m$  receives a link. We call  $f$  the *interaction propensity* with respect to  $m$ . For instance, the classical degree-based PA used in Barabasi-Albert and subsequent models — links attach proportionally to node degrees [2, 3, 8] — is an assumption on  $f$  equivalent to  $f(k) \propto k$ .

$P(m)$  typically denotes the distribution of nodes of type  $m$ . The probability  $P(m|L)$  for a new link extremity to be attached to an agent of kind  $m$  is therefore proportional to  $f(m)P(m)$ , or  $P(L|m)P(m)$ . Applying the Bayes formula yields indeed:<sup>3</sup>

$$P(m|L) = \frac{f(m)P(m)}{P(L)} \quad (1)$$

with  $P(L) = \sum_{m' \in \mathcal{M}} f(m')P(m')$ .

Empirically, during a given period of time  $\nu$  new interactions occur and  $2\nu$  new link extremities appear. Note that a repeated interaction between two already-linked nodes is not considered a new link, for it incurs acquaintance bias. The expectancy of new link extremities attached to nodes of property  $m$  along a period is thus  $\nu(m) = P(m|L) \cdot 2\nu$ . As  $\frac{2\nu}{P(L)}$  is a constant of  $m$  and the network is considered static during the time period, we may estimate  $f$  through  $\hat{f}$  such that:

$$\begin{cases} \hat{f}(m) = \frac{\nu(m)}{P(m)} & \text{if } P(m) > 0 \\ \hat{f}(m) = 0 & \text{if } P(m) = 0 \end{cases} \quad (2)$$

Thus  $\mathbf{1}_P(m)f(m) \propto \hat{f}(m)$ , where  $\mathbf{1}_P(m) = 1$  when  $P(m) > 0$ , 0 otherwise.

## 2.2 Dyadic PA

Adopting a dyadic viewpoint is required whenever a property has no meaning for a single node, which is mostly the case for properties such as proximity, similarity — or distances in general. We therefore intend to measure interaction propensity for a dyad of agents which fulfills a given property  $d$

---

<sup>3</sup>For consistency purposes, we also assume  $f$  strictly positive:  $\forall m \in \mathcal{M}, f(m) > 0$ .

taking values in  $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$ . Similarly, we assume the existence of an essential dyadic interaction behavior embedded into  $g$ , a strictly positive function of  $d$ ; correspondingly the conditional probability  $P(L|d)$ . Again, interaction of a dyad satisfying property  $d$  is  $g(d)$  times more probable. In this respect, the probability for a link to appear between two such agents is:

$$P(d|L) = \frac{g(d)P(d)}{P(L)} \quad (3)$$

with  $P(L) = \sum_{d' \in \mathcal{D}} g(d')P(d')$ .

Here, the expectancy of new links between dyads of kind  $d$  is  $\nu(d) = P(d|L)\nu$ . Since  $\frac{\nu}{P(L)}$  is a constant of  $d$  we may estimate  $g$  with  $\hat{g}$ :

$$\begin{cases} \hat{g}(d) = \frac{\nu(d)}{P(d)} & \text{if } P(d) > 0 \\ \hat{g}(d) = 0 & \text{if } P(d) = 0 \end{cases} \quad (4)$$

Likewise, we have  $\mathbf{1}_P(d)g(d) \propto \hat{g}(d)$ .

## 3 Interpreting interaction propensions

### 3.1 Shaping hypotheses

The PA behavior embedded in  $\hat{f}$  (or  $\hat{g}$ ) for a given monadic (or dyadic) property can be reintroduced as such in modeling assumptions, either (i) by reusing the exact empirically calculated function, or (ii) by stylizing the trend of  $\hat{f}$  (or  $\hat{g}$ ) and approximating  $f$  (or  $g$ ) by more regular functions, thus making possible analytic solutions.

Still, an acute precision when carrying this step is often critical, for a slight modification in the hypotheses (e.g. non-linearity instead of linearity) makes some models unsolvable or strongly shakes up their conclusions. For this reason, when considering a property for which there is an underlying natural order, it may also be useful to examine the cumulative propensity  $\hat{F}(m_i) = \sum_{m'=m_1}^{m_i} \hat{f}(m')$  as an estimation of the integral of  $f$ , especially when the data are noisy (the same goes with  $\hat{G}$  and  $\hat{g}$ ).

### 3.2 Correlations between properties

Besides, if modellers want to consider PA with respect to a collection of properties, they have to make sure that the properties are uncorrelated or that they take into account the correlation between properties: evidence suggests indeed that for instance node degrees depend on age. Often models assume properties to be uncorrelated which, when it is not the case, would amount to count twice a similar effect.<sup>4</sup>

If two distinct properties  $p$  and  $p'$  are independent, the distribution of nodes of kind  $p$  in the subset of nodes of kind  $p'$  does not depend on  $p'$ , i.e. the quantity  $\frac{P(p|p')}{P(p)}$  must theoretically be equal to 1,  $\forall p, \forall p'$ . Empirically, it is possible to estimate it through:<sup>5</sup>

$$\begin{cases} \widehat{c_{p'}}(p) = \frac{P(p|p')}{P(p)} & \text{if } P(p) > 0 \\ \widehat{c_{p'}}(p) = 0 & \text{if } P(p) = 0 \end{cases} \quad (5)$$

in the same manner as previously.

### 3.3 Essential behavior

As such, calculated propensions do not depend on the distribution of nodes of a given type at a given time. In other words, if for example physicists prefer to interact twice more with physicists than with sociologists but there are three times more sociologists around, physicists may well be apparently interacting more with sociologists. Nevertheless,  $\widehat{f}$  remains free of such biases and yields the “baseline” preferential interaction behavior of physicists.

However,  $\widehat{f}$  could still depend on global network properties, e.g. its size, or its average shortest path length. Validating the assumption that  $\widehat{f}$  is independent of *any* global property of the network (notably its topological structure) — i.e., that it is an *entirely essential* property of nodes of kind  $p$  — would require to compare different values of  $\widehat{f}$  for various periods and network configurations. Put differently, this entails checking whether the shape of  $\widehat{f}$  itself is a function of global network parameters.

<sup>4</sup>Like for instance in [16] where effects related to degree and common acquaintances are combined in an independent way.

<sup>5</sup>For computing the correlation between a monadic and a dyadic property, it is easy to interpret  $P(p|d)$  as the distribution of  $p$ -nodes being part of a dyad  $d$ .

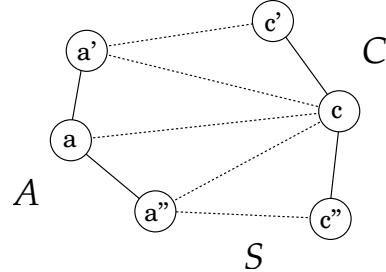


Figure 1: Sample socio-semantic network (3 agents  $a, a', a''$  and 3 concepts  $c, c', c''$ ).

### 3.4 Activity

Additionally,  $\widehat{f}$  represents equivalently an attractivity or an activity: if interactions occur preferentially with some kinds of agents, it could as well mean that these agents are more attractive or that they are more active. If more attractive, the agent will be interacting more, thus being apparently more active. To distinguish between the two effects, it is sometimes possible to measure independently agent activity, notably when interactions occur during *events*, or when interaction initiatives are traceable (e.g. in a directed network).

In such cases, the distinction is far from neutral for modeling. Indeed, when considering evolution mechanisms focused not on agents creating links, but instead on events gathering agents (like in [22]), modellers have to be careful when integrating back into models the observed PA as a behavioral hypothesis. Some categories of agents might in fact be more active and accordingly involved in more events, instead of enjoying more attractivity.

## 4 An application to socio-semantic networks

### 4.1 Definitions

We now apply the above tools to a socio-semantic network, that is, a social network where agents are also linked to semantic items. We examine therein two particular kinds of PA: (i) PA related to a monadic property: the node degree; and (ii) PA linked to a dyadic property: homophily, i.e. the propensity of individuals to interact more with similar agents.

**Networks.** The social network  $\mathcal{A}$  is the network of agents, where links correspond to interactions:  $\mathcal{A} = (\mathbf{A}, E_{\mathbf{A}})$ , with  $\mathbf{A}$  denoting the agent set and  $E_{\mathbf{A}}$  the (undirected) set of links between agents. Interactions occur through events, and each event is associated with a semantic content, made of semantic markers (e.g. keywords), or *concepts*. Dually, the concepts form a semantic network, where concepts are linked to each other if they jointly appear in an event. Identically to  $\mathcal{A}$ , we have  $\mathcal{C} = (\mathbf{C}, E_{\mathbf{C}})$ . Finally, agents are linked to concepts associated with events they are involved in. We hence deal with a third network,  $\mathcal{S} = (\mathbf{A}, \mathbf{C}, E_{\mathbf{AC}})$ , and three kinds of links: (i) between pairs of agents, (ii) between pairs of concepts, and (iii) between concepts and agents. Since we measure agent behavior through network dynamics, we also consider the temporal series of networks  $\mathcal{A}(t)$ ,  $\mathcal{C}(t)$  and  $\mathcal{S}(t)$ , with  $t \in \mathbb{N}$ , which altogether make a dynamic socio-semantic network (see Fig. 1).

**Empirical protocol.** Empirical data come from the bibliographical database Medline which contains dated abstracts of published articles of biology and/or medicine. We focused on a portion concerning a well-defined community of embryologists working on the *zebrafish*, during the period 1997-2004. Translated in the above framework, articles are events, their authors are the agents, and semantic markers are made of expert-selected abstract words.

In order to have a non-empty and statistically significant network for computing propensions, we first build the network on an initialization period of 7 years (from 1997 to end-2003), then carry the calculation on new links appearing during the last year. The dataset contains around 10,000 authors, 5,000 articles and 70 concepts.

## 4.2 Linear degree-related PA

We use Eq. 2 and consider the node degree  $k$  as property  $m$  (thus  $\mathcal{M} = \mathbb{N}$ ): in this manner, we intend to compute the real slope  $\hat{f}(k)$  of the degree-related PA and compare it with the assumption “ $f(k) \propto k$ ”. This hypothesis classically relates to the preferential linking of new nodes to old nodes. To ease the comparison, we considered the subset of interactions between a new and an old node.

Empirical results are shown on Fig. 2. Seem-

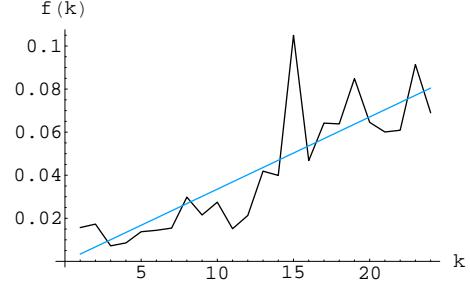


Figure 2: Degree-related interaction propensity  $\hat{f}$ , computed on a one-year period, for  $k < 25$ . The solid line represents the best linear fit.

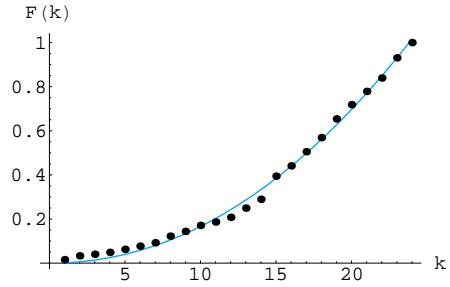


Figure 3: Cumulated propensity  $\hat{F}$ . Dots represent empirical values, solid color lines are the best non-linear fit for  $\hat{F} \sim k^{2.07}$ .

ingly, the best linear fit corroborates the data and tends to confirm that  $f(k) \propto k$ . The best non-linear fit however deviates from this hypothesis, suggesting that  $f(k) \propto k^{1.09}$ . As suggested above, knowing precisely the exponent may be critical here. Since there is a natural order on  $k$ , we plotted the cumulated propensity  $\hat{F}(k) = \sum_{k'=1}^k \hat{f}(k')$  on Fig. 3. In this case, the best non-linear fit for  $\hat{F}$  is  $\hat{F}(k) \propto k^{2.07}$ , confirming the slight deviation from a strictly linear preference which would yield  $k^2$ .

**Rich-work-harder.** This precise result is not new and agrees with existing studies of the degree-related PA (e.g. [15, 19]). Nevertheless, we wish to stress a more fundamental point in computing this kind of PA. Indeed, considerations on agent activity lead us to question the usual underpinnings and justifications of PA related to a monadic property. Concerning degree-related PA, it is the “rich-get-richer” metaphor describing

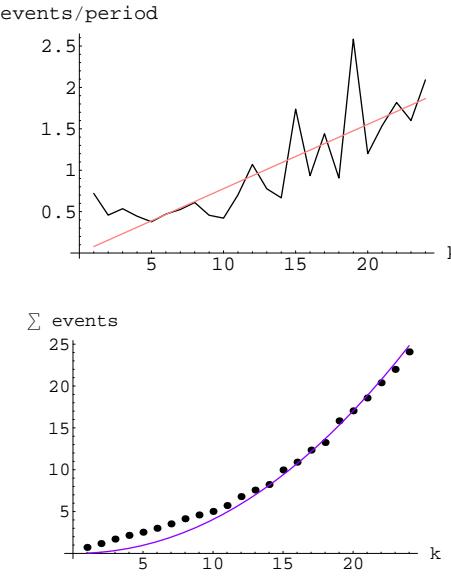


Figure 4: *Top:* Activity  $a(k)$  during the same period, in terms of articles per period (events per period) with respect to agent degree. Best nonlinear fit is  $k^{1.02}$ . *Below:* Cumulated activity  $A(k) = \sum_{k'=1}^k a(k)$ .

rich, or well-connected agents as more attractive than poorly connected agents, thus receiving more connections and becoming even more connected.<sup>6</sup>

When considering the activity of agents with respect to  $k$ , that is, the number of events in which they participate (here, the number of articles they co-author), “rich” agents are proportionally more active than “poor” agents (see Fig. 4), and thus obviously encounter more interactions. It might thus well simply be that richer agents work harder, not are more attractive; the underlying behavior linked to preferential interaction being simply “proportional activity”.<sup>7</sup>

While formally equivalent from the viewpoint of PA measurement, the “rich-get-richer” and “rich-work-harder” metaphors are not behaviorally equivalent. One could choose to be blind to

<sup>6</sup> “(...) the probability that a new actor will be cast with an established one is much higher than that the new actor will be cast with other less-known actors” [2].

<sup>7</sup>Indeed, when considering  $k$  as a proxy for agent activity (i.e., a behavioral feature), if the number of coauthors does not depend on  $k$  (which is actually roughly the case in this data) then observing a linear degree-related PA is not surprising.

this phenomenon and keep an interaction propensity proportional to node degree. On the other hand, one could also prefer to consider higher-degree nodes as more active, assuming instead that the number of links per event is degree-independent and that agents do neither *prefer*, nor *decide* to interact with famous, highly connected nodes; a hypothesis suggested by the present empirical results. If both of these conceptions are consistent with the observed quasi-proportional PA, they bear different implications for modeling, as underlined in Sec. 3.4.

More generally, such feature supports the idea that events, not links, are the right level of modeling for social networks — with events reducing in some cases to a dyadic interaction.

### 4.3 Homophilic PA

Homophily translates the fact that agents prefer to interact with other resembling agents. Here, we assess the extent to which agents are “homophilic” by introducing an inter-agent semantic distance. By semantic distance we mean a function of a dyad of nodes that enjoys the following properties: (i) decreasing with the number of shared concepts between the two nodes, (ii) increasing with the number of distinct concepts, (iii) equal to 1 when agents have no concept in common, and to 0 when they are linked to identical concepts.

Given  $(a, a') \in A^2$  and denoting by  $a^\wedge$  the set of concepts  $a$  is linked to, we introduce a semantic distance  $\delta(a, a') \in [0; 1]$  satisfying the previous properties:<sup>8</sup>

$$\delta(a, a') = \frac{|(a^\wedge \setminus a'^\wedge) \cup (a'^\wedge \setminus a^\wedge)|}{|a^\wedge \cup a'^\wedge|}$$

As  $\delta$  takes real values in  $[0, 1]$  we need to discretize  $\delta$ . To this end, we use a uniform partition of  $[0; 1]$  in  $I$  intervals, to which we add the singleton  $\{1\}$ . We thus define a new discrete property  $d$

<sup>8</sup>Written in a more explicit manner, with  $a^\wedge = \{c_1, \dots, c_n, c_{n+1}, \dots, c_{n+p}\}$  and  $a'^\wedge = \{c_1, \dots, c_n, c'_{n+1}, \dots, c'_{n+q}\}$ , we have  $\delta(a, a') = \frac{p+q}{p+q+n}$ ;  $n$  and  $p, q$  representing respectively the number of elements  $a^\wedge$  and  $a'^\wedge$  have in common and have in proper. We also verify that if  $n = 0$  (disjoint sets),  $\delta(a, a') = 1$ ; if  $n \neq 0$ ,  $p = q = 0$  (same sets),  $\delta(a, a) = 0$ ; and if  $a^\wedge \subset a'^\wedge$  (included sets),  $\delta(a, a') = \frac{q}{q+n}$ .

This distance is very classical and is based on the Jaccard coefficient [4]. It is moreover easy though cumbersome to show that  $\delta(\cdot, \cdot)$  is also a metric distance.

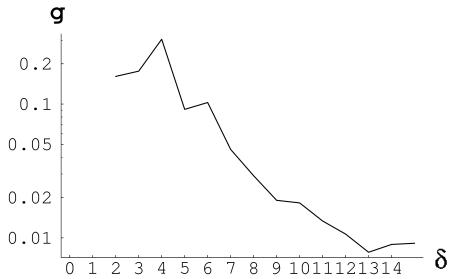


Figure 5: Homophilic interaction propensity  $\hat{g}$  with respect to  $d \in \mathcal{D} = \{d_0, \dots, d_{15}\}$ . The y-axis is in log-scale.

taking values in  $\mathcal{D} = \{d_0, d_1, \dots, d_I\}$  consisting of  $I+1$  intervals:  $\mathcal{D} = \{[0; \frac{1}{I}]; [\frac{1}{I}; \frac{2}{I}]; \dots; [\frac{I-1}{I}; 1]; \{1\}\}$ . Finally, we obtain an empirical estimation of homophily with respect to this distance by applying Eq. 4 on  $d$ , with  $I = 15$ .

The results are gathered on Fig. 5 and show that while agents favor interactions with slightly different agents (as the initial increase suggests), they still very strongly prefer similar agents, as the clearly decreasing trend indicates (sharp decrease from  $d_4$  to  $d_{13}$ , with  $d_4$  being one order of magnitude larger than  $d_{13}$  — note also that  $\hat{g}(d_0) = \hat{g}(d_1) = 0$  because no new link appears for these distance values). Agents thus display semantic homophily, a fact that fiercely advocates the necessity of taking semantic content into account when modeling such social networks.

#### 4.4 Correlation between degree and semantic distance

In other words, the exponential trend of  $\hat{g}$  suggests that scientists seem to choose collaborators most importantly because they are sharing interests, and less because they are attracted to well-connected colleagues, which besides actually seems to reflect agent activity.

As underlined in Sec. 3.1, when building a model of such network based on degree-related and homophilic PA, one has to check whether the two properties are independent, i.e. whether or not a node of low degree is more or less likely to be at a large semantic distance of other nodes. It appears here that there is no correlation between degree and semantic distance: for a given semantic distance  $d$ , the probability of finding a couple of nodes including a node of degree  $k$  is the same as

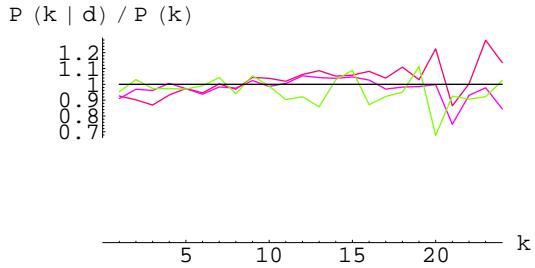


Figure 6: Degree and semantic distance correlation estimated through  $\hat{c}_d(k) = \frac{P(k|d)}{P(k)}$ , plotted here for three different values of  $d$ :  $d \in \{d_5, d_8, d_{11}\}$ .

it is for any value of  $d$  — see Fig. 6. To go further, we suggest that socio-semantic networks might be structured in communities because agents group according to similar interests, in epistemic communities [25], through a mechanism involving events where agents are more or less active, and gather preferentially with respect to their interests; the former being entirely independent of the latter.

## Conclusion

Preferential attachment is the cornerstone of growth mechanisms in most recent social network formation models. This notion was established by the success of a pioneer model [2] rebuilding a major stylized fact of empirical networks, the scale-free degree distribution. While PA has subsequently been widely used, few authors have tried to check or quantify the rather arbitrary assumptions on PA — when such prospects exist, they are mostly dealing with degree-related PA or estimating PA phenomena as single parameters. Models should go further towards empirical investigation when designing hypotheses. This would be really appealing to social scientists, who are usually not seeking *normative* models. We are confident the present reluctance to measuring interaction behaviors and processes is due to the lack of a clean general framework for this purpose, which is the aim of this paper.

We thus introduced the notion of “interaction propensity”, whereby we assume that agents have an *essential* preferential interaction behavior. Using this concept, we designed measurement tools

for quantifying, in a dynamic network, any kind of PA with respect to any property of a single node or of a dyad of nodes — a generalized preferential attachment. The result is a function yielding a comprehensive description of interaction behavior related to a given property. In addition to clarifying PA three new features are crucial: (i) properties not related to the network structure (such as homophily), (ii) correlations between properties, (iii) activity of agents and nature of interactions (e.g. modeling events, not nodes attaching to each other). This kind of hindsight on the notion and status of PA should be useful, even for normative models.

We finally applied these tools to a particular case of socio-semantic network, a scientific collaboration network with agents linked to semantic items. While we restricted ourselves to a reduced example of two significant properties (node degree and semantic distance), measuring PA relatively to other parameters could actually have been very relevant as well — such as PA based on social distance for instance (shortest path length between two agents in the social network). Specifying the list of properties is nevertheless a process driven by the real-world situation *and* by the stylized facts the modeller aims at rebuilding and considers relevant for morphogenesis.

More generally, this framework could be applied to any kind of network, as well as adapted to disconnection propensities. Likewise, once propensities of *interaction in the broad sense* are known, a whole class of social network morphogenesis models [5, 9, 24] can be designed, with agents interacting on a growing network according to stylized interaction heuristics, heuristics precisely based on those measured empirically. *In fine*, introducing more credible hypotheses based on real-case empirical measures would obviously help attract more social scientists in this promising field.

**Acknowledgements.** The author wishes to thank Clémence Magnien, Matthieu Latapy and Paul Bourgine for very fruitful discussions. This work has been partially funded by the CNRS and PERSI.

## References

- [1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.
- [2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [3] A.-L. Barabási, H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, and T. Schubert. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, 2002.
- [4] V. Batagelj and M. Bren. Comparing resemblance measures. *Journal of Classification*, 12(1):73–90, 1995.
- [5] M. Boguna and R. Pastor-Satorras. Class of correlated random networks with hidden variables. *Physical Review E*, 68:036112, 2003.
- [6] M. Boguna, R. Pastor-Satorras, A. Diaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Physical Review E*, 70:056122, 2004.
- [7] G. Caldarelli, A. Capocci, P. D. L. Rios, and M. A. Munoz. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89(25):258702, 2002.
- [8] M. Catanzaro, G. Caldarelli, and L. Pietronero. Assortative model for social networks. *Physical Review E*, 70:037101, 2004.
- [9] P. Cohendet, A. Kirman, and J.-B. Zimmermann. Emergence, formation et dynamique des réseaux – modèles de la morphogenèse. *Revue d’Economie Industrielle*, 103(2-3):15–42, 2003.
- [10] V. Colizza, J. R. Banavar, A. Maritan, and A. Rinaldo. Network structures from selection principles. *Physical Review Letters*, 92(19):198701, 2004.
- [11] E. Eisenberg and E. Y. Levanon. Preferential attachment in the protein network evolution. *Physical Review Letters*, 91(13):138701, 2003.
- [12] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.
- [13] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *ICALP ’02: Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, pages 110–122, London, UK, 2002. Springer-Verlag.
- [14] J.-L. Guillaume and M. Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90(5):215–221, 2004.
- [15] H. Jeong, Z. Néda, and A.-L. Barabási. Measuring preferential attachment for evolving networks. *Europhysics Letters*, 61(4):567–572, 2003.

- [16] E. M. Jin, M. Girvan, and M. E. J. Newman. The structure of growing social networks. *Physical Review E*, 64(4):046132, 2001.
- [17] M. Latapy, C. Magnien, M. Mariadassou, and C. Roth. A basic toolbox for the analysis of dynamics of growing networks. Forthcoming.
- [18] M. McPherson and L. Smith-Lovin. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–440, 2001.
- [19] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review Letters E*, 64(025102), 2001.
- [20] M. E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001.
- [21] W. W. Powell, D. R. White, K. W. Koput, and J. Owen-Smith. Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology*, 110(4):1132–1205, 2005.
- [22] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras. Self-organization of collaboration networks. *Physical Review E*, 70:036106, 2004.
- [23] S. Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58:49–54, 2005.
- [24] C. Roth and P. Bourgine. Binding social and cultural networks: a model. arXiv.org e-print archive, nlin.AO/0309035, 2003.
- [25] C. Roth and P. Bourgine. Epistemic communities: Description and hierachic categorization. *Mathematical Population Studies*, 12(2):107–130, 2005.
- [26] B. Skyrms and R. Pemantle. A dynamic model of social network formation. *PNAS*, 97(16):9340–9346, 2000.
- [27] T. A. Snijders. The statistical evaluation of social networks dynamics. *Sociological Methodology*, 31:361–395, 2001.
- [28] B. Söderberg. A general formalism for inhomogeneous random graphs. *Physical Review E*, 68:026107, 2003.
- [29] H. Stefancic and V. Zlatic. Preferential attachment with information filtering–node degree probability distribution properties. *Physica A*, 350(2-4):657–670, 2005.